A GENERALIZED ENVIRONMENTAL DATA MODEL[*]

Gerald S. Key
Computer Sciences Corporation
San Diego, California 92110-5164 USA
Internet mail: key@cscnet.com
World Wide Web: http://environ.nosc.mil

ABSTRACT

Many environmental processes have long time scales and broad spatial extents. To understand processes on these scales, investigators often must rely on measurements made by others for dissimilar purposes. To do so requires fully documented, primary measurement data stored in a digital format that does not impose a particular view of the data. This paper describes efforts to develop a data model for these purposes. The model has been used to design environmental databases, data reporting specifications, and distributed network architectures. This paper describes the model and discusses its application.

1.0 INTRODUCTION

The U.S. Environmental Protection Agency estimated in 1994 that its regulated community was spending $5 billion annually to perform environmental analyses (USEPA, 1994). This estimate includes only regulatory measurements and focuses primary on analytical chemistry, and therefore probably underestimates the actual costs. Whatever the figure, the real cost is higher still because most of these measurements are used only once, for their original purpose, and then effectively discarded. Personal observations suggest that environmental measurement data have a “half-life” of about 2-3 years. After that period enough information about the measurements (i.e., “metadata”) will have been lost to render half of the measurements unusable for other purposes. Stated another way, nearly all the measurements will be unusable after 10 years. Coincidentally, many of processes of interest to environmental scientists have turnover rates on the order of decades. Thus, even if we could afford to make the measurements again, we still will not have the 10- to 20-year perspective necessary to detect significant environmental change. The same argument pertains to the loss of perspective for making spatial, methodological, and a host of other comparisons of these data.

To reuse environmental measurements, the data must be readily accessible and contain the information the new application requires. These two prerequisites, accessibil-

[*] Presented at Eco-Informa ’96, Lake Buena Vista, Florida, 4-7 November 1996.

ity and utility, seem obvious until you ask what information you would record today, in what format and on what medium, to make those measurements useful in10 years for an application that may not exist today? The remainder of this paper develops the argument that the answer to this question is that you must record **primary measurement data in fully documented digital form**.

A **primary measurement** is a quantitative observation made in the field or laboratory. It includes what was measured, the quantity of the measured parameter, and the units in which the quantity is expressed. Examples of primary measurements include:

12.9 mg/L copper
47 *Acanthurus sandvicensis*
3.6 cm/sec water velocity

Means, standard deviations, diversity indices, and other summary statistics cannot substitute for primary measurement data. They explain less of the variance (i.e., loss of information) and their methods of calculation are subject to change.

A **fully documented** primary measurement includes supporting information and associated measurements. Supporting information places the measurement in context: where, when, how, by whom, etc. was the measurement made? Supporting information may have its own additional information requirements. For example, recording the geographic coordinates should also include the reference datum and the significant digits in the coordinate values. Associated measurements are made to specify the quality of the primary measurements. Associated measurements might include duplicate and replicate analyses, the detection limit(s) of the analytical method, etc.

Storing environmental measurements in **digital form** is a prerequisite for several reasons. Key-entering data from a hardcopy source is too time-consuming, expensive, and error-prone in most applications. The physical space required to store 10-20 years of data in hardcopy form is also impractical. Digital media can be made more permanent because of their ability to detect and correct errors on deteriorated media. And finally, most environmental measurements are now recorded digitally or converted to digital form for processing. Retaining these measurements in digital form is frequently easier than converting them to hardcopy.

The Environmental Sciences Division at the Naval Command, Control and Ocean Surveillance Center RDT&E Division (NRaD) in San Diego, California, USA is developing a generalized environmental data model to meet these requirements. NRaD is using this model to implement multi-disciplinary environmental databases, organize and enter historical measurement data into these databases, prepare specifications for reporting environmental measurements, and to share data with other Navy and non-Navy projects.

NRaD is also actively pursuing the expansion of a generalized environmental data model to a national or international scale.

2.0 BACKGROUND

The design requirements for a generalized database of environmental measurements include:

- Primary Measurements. The database must accommodate primary measurements made by different disciplines (e.g., biology, chemistry), in different media (e.g., sediment, water), using different sampling methods (e.g., discrete, continuous).
- Full Documentation. The structure of the database should serve as a template for specifying how to report data from external sources and share them among various users.
- No *a priori* View. The database should accommodate the measurements of different disciplines without imposing the perspectives of those disciplines on the structure of data.
- Distributed Data. Sharing measurement data will, in many instances, involve linking local and regional databases locations into a larger computer network.
- Growth. The design should allow the structure of the database to change without requiring the redesign of the applications that use it. Conversely, users should be able to upgrade their application software without having to redesign the database.

Michener, *et al*. (1994) and others[1] have noted the movement within the environmental sciences toward better documentation of measurement data. However, because most environmental measurement data sets are stored in “flat files”, these efforts have focused on “metadata” standards for documenting the contents of these files[2]. NRaD’s approach has been to integrate data and metadata into a common **data model**.

A data model is the logical representation of an organization’s data (see Simsion, 1994). It specifies what data to store and how to organize them. The most commonly used data modeling technique is the entity-relation (E-R) model. In an E-R model, *entities* represent the real-world objects about which data are to be stored, *relationships* define how one entity relates to another, and *attributes* are the pieces of information recorded about an entity. Key (1996) discusses the application of the E-R methodology to the NRaD environmental data model.

[1] http://www.sdsc.edu/Events/compeco_workshop/master.html A reference which begins “http://” is a Uniform Resource Locator (URL) for a World-Wide Web site on the Internet.

[2] http://geochange.er.usgs.gov/pub/tools/metadata/standard/metadata.html

3.0 RESULTS

The data model developed by NRaD has been implemented using ERWin/ERX® by LogicWorks, Inc. The model currently includes approximately 50 entities and 60 relationships. Figure 1 illustrates a portion of that data model.

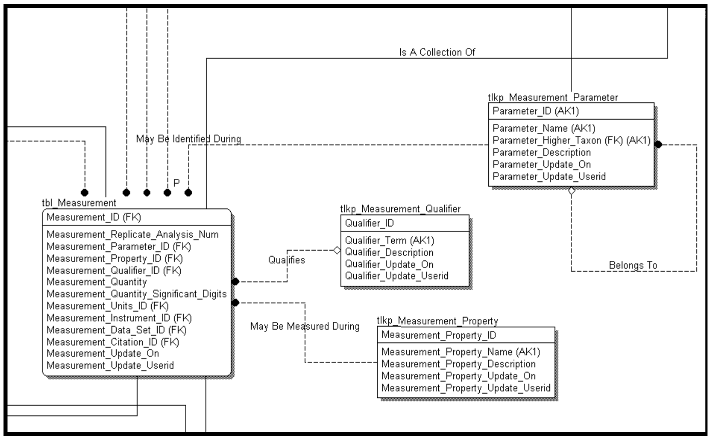
Is A Collection Of
tlkp_Measurement_Parameter
Parameter_ID (AK1)
Parameter_Name (AK1)
Parameter_Higher_Taxon (FK) (AK1)
Parameter_Description
Parameter_Update_On
Parameter_Update_Userid
May Be Identified During
P
tbl_Measurement
Measurement_ID (FK)
Measurement_Replicate_Analysis_Num
Measurement_Parameter_ID (FK)
Measurement_Property_ID (FK)
Measurement_Qualifier_ID (FK)
Measurement_Quantity
Measurement_Quantity_Significant_Digits
Measurement_Units_ID (FK)
Measurement_Instrument_ID (FK)
Measurement_Data_Set_ID (FK)
Measurement_Citation_ID (FK)
Measurement_Update_On
Measurement_Update_Userid
Qualifies
tlkp_Measurement_Qualifier
Qualifier_ID
Qualifier_Term (AK1)
Qualifier_Description
Qualifier_Update_On
Qualifier_Update_Userid
Belongs To
May Be Measured During
tlkp_Measurement_Property
Measurement_Property_ID
Measurement_Property_Name (AK1)
Measurement_Property_Description
Measurement_Property_Update_On
Measurement_Property_Update_Userid

Figure 1. Generalized Environmental Data Model (part)

The central design objective of the NRaD model has been to keep all measurements at the same level of organization. This objective is derived from the requirement that users be able to search the database for measurements without prior knowledge of whether the measurements were made in water, sediment, or tissue, by a particular method, for a specific project, and so on. In the NRaD model, all measurements are stored in the same table (see the **tbl_Measurement** entity in Figure 1), as separate rows (records). Thus, retrieving all measurements of copper is equivalent to retrieving all records from tbl_Measurement where Measurement_Parameter_ID = "Copper"[3], regardless of how or where the measurement was made. The relationships (represented as lines in Figure 1) between tbl_Measurement and other entities enable users to display the measurement medium, method, etc. associated with the copper measurements they have retrieved. Alternatively, the user could used these relationships to select only those measurements of copper made by a particular method, medium, or project if so desired.

The relationship **Belongs_To** in Figure 1 illustrates another common characteristic of a generalized environmental data model: unary, or recursive, relationships. A unary relationship joins an entity to itself. It is used to "flatten out" hierarchical relationships.

[3] In this design the string "Copper" is represented by a surrogate ID value.

In this particular example, the tbl_Measurement_Parameter entity, which identifies the measured parameter (e.g., Copper, *Acanthurus sandvicensis*, Water Velocity), is joined to itself to represent the hierarchical classification systems (i.e., taxonomies) of chemical, biological, and other parameters. Thus, the query Parameter_Higher_Taxon = "Heavy Metals" would produce a set of records with Parameter_Name values such as "Copper", "Lead", "Zinc", etc. This set of records, when joined to the tbl_Measurement table, would yield all the measurements of the chemical species classified as "Heavy Metals". Such recursion can be to any depth: a record for Parameter_Name = *Acanthurus sandvicensis* might point to a record for Parameter_Name = Acanthuridae (the Family to which *Acanthurus* belongs), which in turn points to a record for Parameter_Name = Perciformes (the Order to which Acanthuridae belongs), and so on.

Recursive relationships are important for representing other types of hierarchical relationships, such as samples and subsamples. A tissue sample might be collected from a particular fish, of a particular species, from a particular trawl, on a given cruise, of a specific project. Being able to associate measurements with an event like "trawl" may be as useful for some applications as the ability to retrieve all measurements of a biological or chemical species is to another investigator.

Representing measurement units is another problem area for a generalized data model. When the data model is translated into a database design, performance considerations argue for storing similar measurements in the same units to facilitate relational queries. That is, the query "Measurement_Quantity > 5" will not return the expected results if some measurements are stored in mg/L and others in µg/L. Conversely, normalizing some measurements to common units, like counts of organisms for different sample areas or volume, may introduce unwarranted assumptions and yield unusual quantities. When units are converted, the requirement for full documentation mandates that the original units and the conversion algorithm also be associated with the measurements.

The development of the data model continues to be an evolutionary process. Each new application of the model has expanded the definition of "environmental information," although the core entities relating to measurements, samples, etc. have remained relatively stable since the early versions of the mode. The model has been used to design databases for several projects, including a multi-disciplinary study of sediment contamination near the San Diego Naval Station, an ecological risk assessment in Piscataqua River and Great Bay Estuary, and an integrated compliance and assessment program for the Naval Shipyards. It is also being used to generate specifications for collecting and reporting environmental measurements and for evaluating remote access to environmental data using both conventional client/server and World-Wide Web technology.

4.0 CONCLUSIONS

The environmental data model under development must be tested under a broader range of applications to be considered "generalized." It has been applied to both discrete and continuous sampling regimes, but not to high-volume applications such as remote sensing. It has been used to manage a range of chemical and biological measurements, but to a lesser extent geological and physical data. In addition, there are already indications that relational database management systems may be inadequate for representing the complex data types of environmental measurements. Nonetheless, our testing does indicate the potential for environmental scientists to define a common set of entities, attributes and relationships for documenting environmental measurements. If such a definition could agreed on and adhered to, it would establish a basis for linking local databases into regional or larger aggregations as requirements, technology, and resources permit.

5.0 REFERENCES

Key, Gerald S., "Some Experiences Developing a Generalized Environmental Data Model", *Conference Proceedings, Oceanology International 1996*, Brighton, UK, 1996

Michener, W. K., J. W. Brunt and S. G. Stafford (eds), *Environmental Information Management and Analysis*. Taylor & Francis, London, 1994

Simsion, Graeme, *Data Modeling Essentials*, International Thomson Computer Press, Boston, MA, 1994.

USEPA, *Guidance for the Data Quality Objectives Process*, EPA QA/G-4, Quality Assurance Management Staff, U.S. Environmental Protection Agency, Washington, D.C. 1994

ACKNOWLEDGEMENTS

The author wishes to acknowledge the assistance Marissa Caballero and Andy Patterson of Computer Sciences Corporation for their technical assistance, and Jeff Grovhoug, Ron Gauthier and Bart Chadwick of the Environmental Sciences Division at NRaD for their encouragement and support. This work was supported by Contract Number N66001-94-D-0090 from the Naval Command, Control and Ocean Surveillance Center.